ESCAPE

European Science Cluster of Astronomy &
Particle physics ESFRI research Infrastructures

The Future ESCAPE Data Access & Analysis Framework

John Swinbank — swinbank@astron.nl

ESCAPE - The European Science Cluster of Astronomy & Particle Physics ESFRI Research Infrastructures has received funding from the European Union’s Horizon 2020 research and innovation programme under the Grant Agreement n° 824064.

ESCAPE
European Science Cluster of Astronomy & Particle physics ESFRI research Infrastructures

Goal

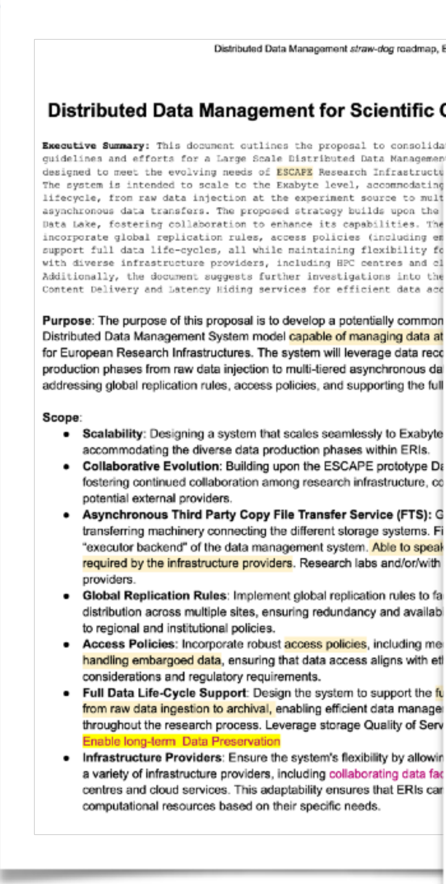
- Unify the ESCAPE Virtual Research Environment (VRE) and Science Analysis Platform (ESAP) efforts behind a coherent roadmap.
- Provide a forum through which ESCAPE Open Collaboration members, and others, can exchange news, views & best practices.
- Provide a modular, scalable analysis platform that RIs can build on and customize to meet their individual need.

Funded by the European Union's
Horizon 2020 - Grant N° 824064

Process

1. Establish a *common vision* 🕶️
2. Define a roadmap to achieve that vision 🗺️
3. Break that roadmap into achievable (fundable?) projects 💰

Vision



Distributed Data Management *straw-dog* roadmap, ESCAPE Collaboration
v0

Distributed Data Management for Scientific Computing

Executive Summary: This document outlines the proposal to consolidate common
guidelines and efforts for a Large Scale Distributed Data Management System
designed to meet the evolving needs of ESCAPE Research Infrastructu
The system is intended to scale to the Exabyte level, accommodating
lifecycle, from raw data injection at the experiment source to mult
asynchronous data transfers. The proposed strategy builds upon the
Data Lake, fostering collaboration to enhance its capabilities. The
incorporate global replication rules, access policies (including em
support full data life-cycles, all while maintaining flexibility fo
with diverse infrastructure providers, including HPC centres and cl
Additionally, the document suggests further investigations into the
Content Delivery and Latency Hiding services for efficient data acc

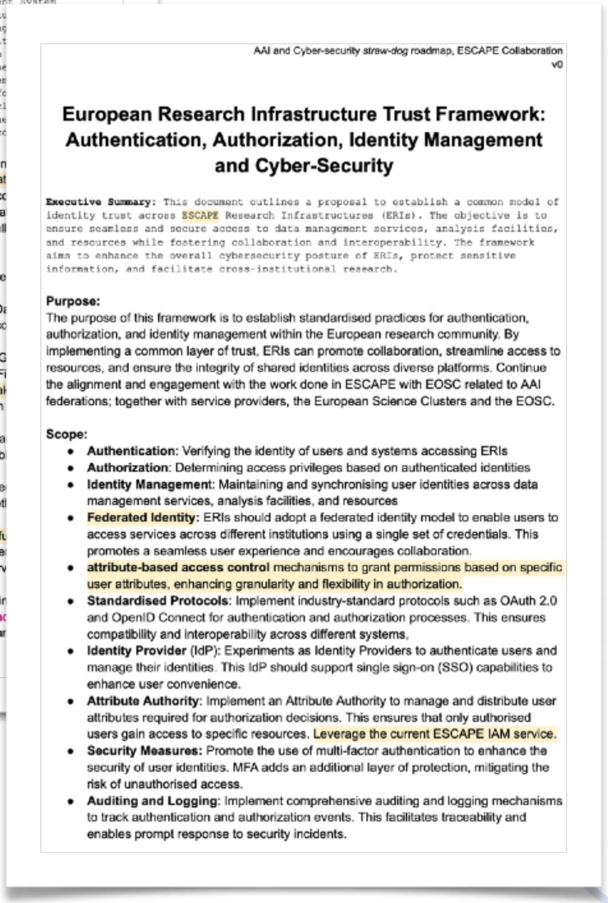
Purpose: The purpose of this proposal is to develop a potentially common
Distributed Data Management System model capable of managing data at
for European Research Infrastructures. The system will leverage data reco
production phases from raw data injection to multi-tiered asynchronous da
addressing global replication rules, access policies, and supporting the full

Scope:

- **Scalability**: Designing a system that scales seamlessly to Exabyte
  accommodating the diverse data production phases within ERIs.
- **Collaborative Evolution**: Building upon the ESCAPE prototype Da
  fostering continued collaboration among research infrastructure, co
  potential external providers.
- **Asynchronous Third Party Copy File Transfer Service (FTS):** G
  transferring machinery connecting the different storage systems. Fi
  "executor backend" of the data management system. Able to speak
  required by the infrastructure providers. Research labs and/or/with
  providers.
- **Global Replication Rules**: Implement global replication rules to fa
  distribution across multiple sites, ensuring redundancy and availab
  to regional and institutional policies.
- **Access Policies**: Incorporate robust access policies, including me
  handling embargoed data, ensuring that data access aligns with et
  considerations and regulatory requirements.
- **Full Data Life-Cycle Support**: Design the system to support the fu
  from raw data ingestion to archival, enabling efficient data manage
  throughout the research process. Leverage storage Quality of Serv
  Enable long-term Data Preservation
- **Infrastructure Providers**: Ensure the system's flexibility by allowin
  a variety of infrastructure providers, including collaborating data fac
  centres and cloud services. This adaptability ensures that ERIs car
  computational resources based on their specific needs.

https://edu.nl/c9eyc

https://edu.nl/prphe



AAI and Cyber-security *straw-dog* roadmap, ESCAPE Collaboration
v0

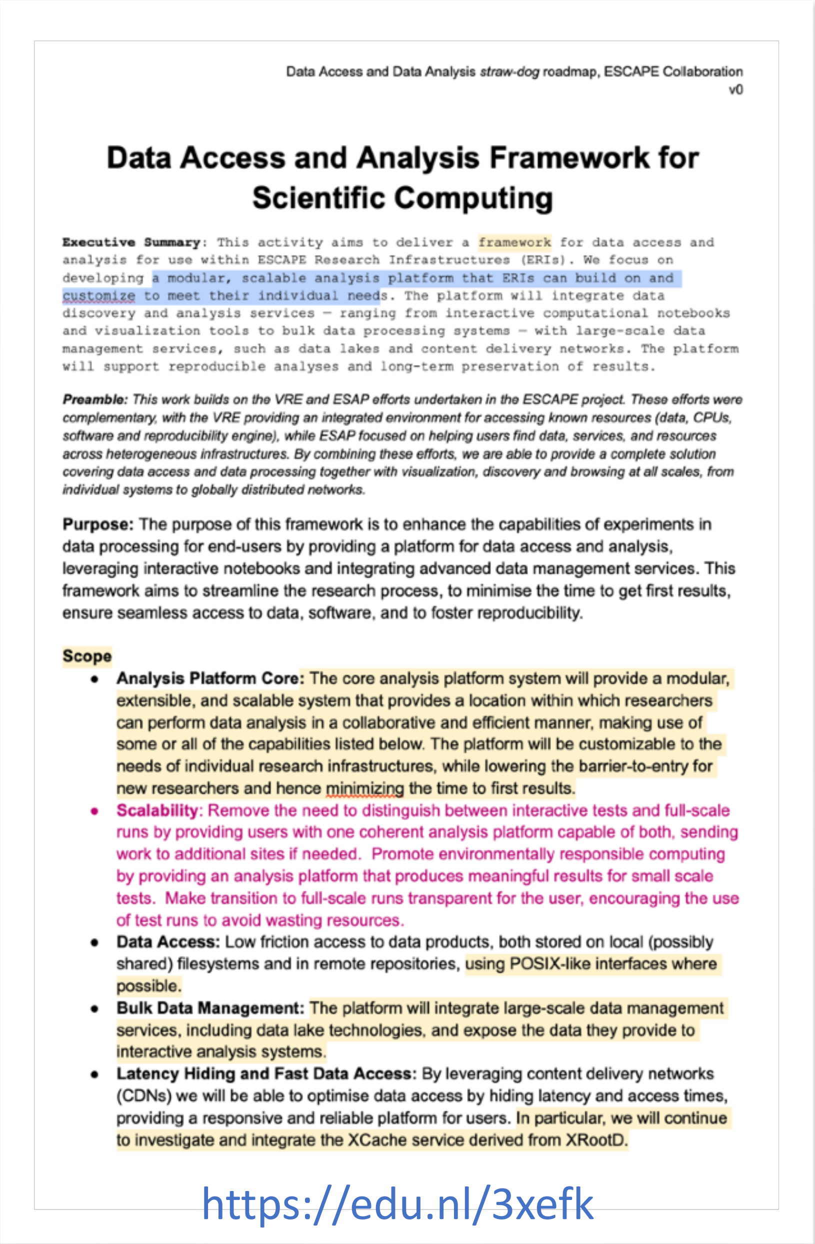
European Research Infrastructure Trust Framework: Authentication, Authorization, Identity Management and Cyber-Security

Executive Summary: This document outlines a proposal to establish a common model of identity trust across ESCAPE Research Infrastructures (ERIs). The objective is to ensure seamless and secure access to data management services, analysis facilities, and resources while fostering collaboration and interoperability. The framework aims to enhance the overall cybersecurity posture of ERIs, protect sensitive information, and facilitate cross-institutional research.

Purpose:
The purpose of this framework is to establish standardised practices for authentication, authorization, and identity management within the European research community. By implementing a common layer of trust, ERIs can promote collaboration, streamline access to resources, and ensure the integrity of shared identities across diverse platforms. Continue the alignment and engagement with the work done in ESCAPE with EOSC related to AAI federations; together with service providers, the European Science Clusters and the EOSC.

Scope:

- **Authentication**: Verifying the identity of users and systems accessing ERIs
- **Authorization**: Determining access privileges based on authenticated identities
- **Identity Management**: Maintaining and synchronising user identities across data management services, analysis facilities, and resources
- **Federated Identity**: ERIs should adopt a federated identity model to enable users to access services across different institutions using a single set of credentials. This promotes a seamless user experience and encourages collaboration.
- **attribute-based access control** mechanisms to grant permissions based on specific user attributes, enhancing granularity and flexibility in authorization.
- **Standardised Protocols**: Implement industry-standard protocols such as OAuth 2.0 and OpenID Connect for authentication and authorization processes. This ensures compatibility and interoperability across different systems.
- **Identity Provider** (IdP): Experiments as Identity Providers to authenticate users and manage their identities. This IdP should support single sign-on (SSO) capabilities to enhance user convenience.
- **Attribute Authority**: Implement an Attribute Authority to manage and distribute user attributes required for authorization decisions. This ensures that only authorised users gain access to specific resources. Leverage the current ESCAPE IAM service.
- **Security Measures:** Promote the use of multi-factor authentication to enhance the security of user identities. MFA adds an additional layer of protection, mitigating the risk of unauthorised access.
- **Auditing and Logging**: Implement comprehensive auditing and logging mechanisms to track authentication and authorization events. This facilitates traceability and enables prompt response to security incidents.



Data Access and Data Analysis *straw-dog* roadmap, ESCAPE Collaboration
v0

Data Access and Analysis Framework for Scientific Computing

Executive Summary: This activity aims to deliver a framework for data access and analysis for use within ESCAPE Research Infrastructures (ERIs). We focus on developing a modular, scalable analysis platform that ERIs can build on and customize to meet their individual needs. The platform will integrate data discovery and analysis services — ranging from interactive computational notebooks and visualization tools to bulk data processing systems — with large-scale data management services, such as data lakes and content delivery networks. The platform will support reproducible analyses and long-term preservation of results.

Preamble: *This work builds on the VRE and ESAP efforts undertaken in the ESCAPE project. These efforts were complementary, with the VRE providing an integrated environment for accessing known resources (data, CPUs, software and reproducibility engine), while ESAP focused on helping users find data, services, and resources across heterogeneous infrastructures. By combining these efforts, we are able to provide a complete solution covering data access and data processing together with visualization, discovery and browsing at all scales, from individual systems to globally distributed networks.*

Purpose: The purpose of this framework is to enhance the capabilities of experiments in data processing for end-users by providing a platform for data access and analysis, leveraging interactive notebooks and integrating advanced data management services. This framework aims to streamline the research process, to minimise the time to get first results, ensure seamless access to data, software, and to foster reproducibility.

Scope

- **Analysis Platform Core:** The core analysis platform system will provide a modular, extensible, and scalable system that provides a location within which researchers can perform data analysis in a collaborative and efficient manner, making use of some or all of the capabilities listed below. The platform will be customizable to the needs of individual research infrastructures, while lowering the barrier-to-entry for new researchers and hence minimizing the time to first results.
- **Scalability**: Remove the need to distinguish between interactive tests and full-scale runs by providing users with one coherent analysis platform capable of both, sending work to additional sites if needed. Promote environmentally responsible computing by providing an analysis platform that produces meaningful results for small scale tests. Make transition to full-scale runs transparent for the user, encouraging the use of test runs to avoid wasting resources.
- **Data Access:** Low friction access to data products, both stored on local (possibly shared) filesystems and in remote repositories, using POSIX-like interfaces where possible.
- **Bulk Data Management:** The platform will integrate large-scale data management services, including data lake technologies, and expose the data they provide to interactive analysis systems.
- **Latency Hiding and Fast Data Access:** By leveraging content delivery networks (CDNs) we will be able to optimise data access by hiding latency and access times, providing a responsive and reliable platform for users. In particular, we will continue to investigate and integrate the XCache service derived from XRootD.

https://edu.nl/3xefk

Funded by the European Union's
Horizon 2020 - Grant N° 824064

Roadmap

- https://strawpoll.com/05ZdW8VmNg6

Funded by the European Union's
Horizon 2020 - Grant N° 824064

Funding

- OSCARS cascading grant.
  - First call expected March 2024.
  - Projects up to €250k.
  - Future call towards the end of this year.
- Other projects & opportunities will likely also exist.

OSCARS
Open Science Clusters' Action
for Research & Society

Funded by the European Union's
Horizon 2020 - Grant N° 824064

Admin

- John will be scaling back his involvement
  - Still keen to participate in discussions and help define the roadmap!
  - But it's time to step back from chairing this group.
- Enrique will coordinate the development of the roadmap.
- We're looking for a volunteer from the ESAP community to step up and help Enrique out.

Funded by the European Union's Horizon 2020 - Grant N° 824064